36-303: Sampling, Surveys and Society
Exam 2
Tue Apr 17, 2012

- You have 80 minutes for this exam.
- The exam is closed-book, closed notes.
- A calculator is allowed.
- **Two formula sheets are provided for your convenience.**
- Please write all your answers on the exam itself; your work must be your own.
- If you need more room, continue onto the back of the same page as the question you are answering *(and let us know that is what you are doing!)*.

Question	Points Possible	Points Earned
1	20	
2	24	
3	18	
4	18	
5	20	
Total	100	

Name: ______________________________

Signature: ______________________________

Some Useful Formulas From the Statistics of Survey Sampling, I

Equally-Likely Outcomes & Counting

- If K outcomes $O_1, \ldots, O_K$ are equally likely, then the probability of any one of them is $1/K$.
- Consider taking a sample of n objects from a population of N objects.
  - Sampling with replacement, there are N^n possible samples of size n; the probability of any one of them is $1/N^n$.
  - Sampling without replacement, there are $\binom{N}{n} = \frac{N!}{n!(N-n)!}$ possible samples of size n [where $N! = N \cdot (N-1) \cdot (N-2) \cdots 3 \cdot 2 \cdot 1$], so the probability of any one of them is $1 \Big/ \binom{N}{n}$.

Discrete Random Variables

Let X and Y be random variables with sample spaces $\{x_1, \ldots, x_K\}$ and $\{y_1, \ldots, y_K\}$ and distributions

$$P[X = x_i, Y = y_j] \;=\; p_{ij} \;\;,\;\; P[X = x_i] \;=\; p_{i\cdot} \;=\; \sum_{j=1}^{K} p_{ij} \;\;,\;\; P[Y = y_j] \;=\; p_{\cdot j} \;=\; \sum_{i=1}^{K} p_{ij}$$

Then, for example

$$E[X] \;=\; \sum_{i=1}^{K} x_i p_{i\cdot} \;\;,\;\; Var(X) \;=\; \sum_{i=1}^{K} (x_i - E[X])^2 p_{i\cdot} \;\;,\;\; Cov(X,Y) \;=\; \sum_{i=1}^{K} (x_i - E[X])(y_i - E[Y]) p_{ij}$$

$$P[X = x_i | Y = y_j] \;=\; p_{ij}/p_{\cdot j} \;\;,\;\; E[X|Y = y_j] \;=\; \sum_{i=1}^{K} x_i P[X = x_i | Y = y_j] \;\;,\;\; E[aX + bY + c] \;=\; aE[X] + bE[Y] + c$$

Random Sampling From a Finite Population

Consider a population of size N and a sample of size n. Let y_i be the (fixed) values of some variable of interest in the population (such as a person's age, or whether they would vote for Obama). Let

$$Z_i = \begin{cases} 1, & \text{if } i \text{ is in the sample} \\ 0, & \text{else} \end{cases}$$

be the random sample inclusion indicators, and let Y_i be the random observations in the sample. Then the sample average can be written

$$\overline{Y} \;=\; \frac{1}{n} \sum_{i=1}^{n} Y_i \;=\; \frac{1}{n} \sum_{i=1}^{N} Z_i y_i$$

The Z_i's are Bernoulli random variables with

$$E[Z_i] \;=\; \frac{n}{N} \;\;,\;\; Var(Z_i) = \frac{n}{N}\left(1 - \frac{n}{N}\right) \;\;,\;\; Cov(Z_i, Z_j) = -\frac{1}{N-1}\frac{n}{N}\left(1 - \frac{n}{N}\right)$$

Confidence Intervals and Sample Size

(a) A CLT-based $100(1-\alpha)\%$ confidence interval for the population mean is $(\overline{Y} - z_{\alpha/2} SE \;,\; \overline{Y} + z_{\alpha/2} SE)$.

(b) For sampling with replacement from an infinite population, $SE = SD/\sqrt{n}$.

(c) For sampling without replacement from a finite population, the SE has to be multiplied by the finite population correction (FPC).

(d) For a given margin of error (ME, half the width of the CI) and confidence level $1-\alpha$, we can find the sample size by solving

$$z_{\alpha/2} SE < ME$$

for n. The same approach works for both SRS with replacement (using the SE in (b)) and SRS without replacement (using the SE in (c)).

Some Useful Formulas From the Statistics of Survey Sampling, II

Stratified Sampling

Consider H strata with population counts $N = \sum_{h=1}^{H} N_h$ and sample counts $n = \sum_{h=1}^{H} n_h$. Let $f_h = n_h/N_h$; $W_h = N_h/N$; and $\overline{y}_h = \frac{1}{n_h}\sum_{i=1}^{n_h} y_{ih}$ in each stratum, and let $s_h^2 = \frac{1}{n_h-1}\sum_i (y_{ih} - \overline{y}_h)^2$ be the sample variance in each stratum. Then

$$\overline{y}_{st} = \sum_{h=1}^{H} W_h \overline{y}_h \ , \quad \mathrm{Var}\,(\overline{y}_{st}) \approx \sum_{h=1}^{H} W_h^2 (1-f_h)\frac{s_h^2}{n_h} \ , \quad DEFF = \frac{\mathrm{Var}\,(\overline{y}_{st})}{\mathrm{Var}\,(\overline{y}_{srs})} = \frac{\sum_{h=1}^{H} W_h^2(1-f_h)\frac{s_h^2}{n_h}}{(1-f)\frac{s^2}{n}}$$

Cluster Sampling

Consider a population of N clusters. We take an SRS $\mathcal{S}$ of n clusters, and all units within each sampled cluster (one-stage clustering). Assume clusters all have same size M. Let $\overline{y}_i = \frac{1}{M}\sum_{j=1}^{M} y_{ij}$ in each cluster. Then

$$\overline{y}_{cl} = \frac{1}{n}\sum_{i\in\mathcal{S}} \overline{y}_i \ , \quad \mathrm{Var}\,(\overline{y}_{cl}) \approx \left(1-\frac{n}{N}\right)\frac{1}{n}s_{\overline{y}_i}^2 = \left(1-\frac{n}{N}\right)\frac{1}{n}\left[\frac{1}{n-1}\sum_{i\in\mathcal{S}}(\overline{y}_i - \overline{y}_{cl})^2\right]$$

and

$$DEFF = \frac{\mathrm{Var}\,(\overline{y}_{cl})}{\mathrm{Var}\,(\overline{y}_{srs})} = \frac{M s_{\overline{y}_i}^2}{s_{y_{ij}}^2} \approx 1 + (M-1)\rho$$

where $s_{\overline{y}_i}^2$ is the sample varance of the cluster means, $s_{y_{ij}}^2$ is the sample variance of the individual observations, and ρ is the intraclass (intracluster) correlation, or ICC.

Post-Stratification Weights and Means

As part of survey data collection it is a good idea to get general demographic information (e.g. in our surveys: sex, age, class, major, hometown, etc.). After data collection we compare the proportions in each of these categories in our sample with the same proportions in the population. If they agree, great. If not, calculate

$$w_i = (N_h/N)/(n_h/n) \text{ for each } i \text{ in post-stratum } h \ , \quad \textit{and} \quad \overline{y}_w = \frac{\sum_i w_i y_i}{\sum_i w_i}$$

Post-Stratification Variance Calculations

<u>*Taylor series*</u>:

$$\mathrm{Var}_{TS}\,(\overline{y}_w) \approx \frac{1}{(\sum_i w_i)^2}\left[\mathrm{Var}\,(\sum_i w_i y_i) - 2\overline{y}_w \mathrm{Cov}\,(\sum_i w_i y_i, \sum_i w_i) + (\overline{y}_w)^2 \mathrm{Var}\,(\sum_i w_i)\right]$$

where $\overline{y}_w$ is as above, $\overline{w} = \frac{1}{n}\sum_i w_i$, $\overline{wy} = \frac{1}{n}\sum_i w_i y_i$,

$$\mathrm{Var}\,(\textstyle\sum_{i=1}^n w_i) \approx n\cdot\frac{1}{n-1}\sum_{i=1}^n (w_i - \overline{w})^2, \quad \mathrm{Var}\,(\sum_{i=1}^n y_i w_i) \approx n\cdot\frac{1}{n-1}\sum_{i=1}^n (w_i y_i - \overline{wy})^2,$$

$$\mathrm{Cov}\,(\textstyle\sum_{i=1}^n y_i w_i, \sum_{i=1}^n w_i) \approx n\cdot\frac{1}{n-1}\sum_{i=1}^n (w_i y_i - \overline{wy})(w_i - \overline{w})$$

<u>*Jackknife*</u>:

- Replicate n times (by removing one obs. each time and recalculating weights):

$$\overline{y}_w^{(r)} = \frac{\sum_{i=1}^n w_i^{(r)} y_i^{(r)}}{\sum_{i=1}^n w_i^{(r)}}$$

- Calculate

$$\overline{y}_{JK} = \frac{1}{n}\sum_{r=1}^{n} \overline{y}_w^{(r)} \ , \quad Var_{JK}(\overline{y}_w) \approx \frac{n-1}{n}\sum_{r=1}^{n}(\overline{y}_w^{(r)} - \overline{y}_{jk})^2$$

1. [20 pts] *Multiple Choice (4 parts). For each part, circle the roman numeral of the one best answer.*

   (a) [5 pts] Let Y_i be the number of monthly neighborhood watch meetings attended by the i^{th} resident in a neighborhood, in the last year. You are going to conduct a survey, using a SRS to estimate $\overline{Y}_{pop}$, the population mean number of meetings attended by neighborhood residents. Among the N_R residents who *would respond* to your survey, the mean number of meetings is $\overline{Y}_R$, and among the N_M number of residents who would *not respond*, the mean number of meetings attended is $\overline{Y}_M$. In class we showed that the *bias* between $\overline{Y}_R$ and $\overline{Y}_{pop}$, due to missing responses, is

   $$\overline{Y}_R - \overline{Y}_{pop} = \frac{N_M}{N}(\overline{Y}_R - \overline{Y}_M)$$

   where $N = N_R + N_M$.

   Which statement below is false (or, circle iv. if all are OK)?

   i. The more people who respond to the survey, the smaller the bias due to missing responses.
   ii. The bigger the difference between mean number of meetings attended by nonresponders, vs the mean number attended by responders, the bigger the bias due to missing responses.
   iii. The larger your SRS, the better you can estimate $\overline{Y}_{POP}$.
   iv. All of the above statements are true.

   (b) [5 pts] Suppose we divide a sampling frame into groups, which we may treat as either strata for stratified sampling, or clusters for cluster sampling. If we make the groups so that *observations* **within groups** *are more* **similar** *to each other*, and *observations* **between groups** *are more* **different** *from each other*, then, all other things being equal, we expect

   i. The variance of the stratified sample mean $\overline{y}_{st}$ will go **down** and the variance of the cluster sample mean $\overline{y}_{cl}$ will go **up**.
   ii. The variance of the stratified sample mean $\overline{y}_{st}$ will go **up** and the variance of the cluster sample mean $\overline{y}_{cl}$ will go **down**.
   iii. Both variances will go **down**.
   iv. Both variances will go **up**.

[Continued on next page...]

(c) [5 pts] In one-stage clustered sampling, the ICC ρ measures

i. The correlation between observations in different clusters.

ii. The correlation between the cluster means of different clusters.

iii. The correlation between observations in the same cluster.

iv. The correlation between the cluster mean and the individual observations in the cluster.

(d) [5 pts] Which of the following is *not* a usual part of post-survey processing?

i. Coding short-answer or text data

ii. Variance calculation

iii. Imputation

iv. Checking post-strata and building weights if needed

v. All of the above *are* usually part of post-survey processing!

[This space intentionally left blank]

2. [24 pts] Many universities in the United States stay open during winter break; closing during break would save some salary and maintenance costs, but would be inconvenient for faculty and staff who want to continue working while students are gone. In January 1995, the Office of University Evaluation at Arizon State University surveys faculty and staff members to find out their reaction the closure of the university during Winter Break 1994. Four strata were identified, and surveys were distributed to an SRS (without replacement) in each stratum. The table below gives the results for one of the survey questions, "*Would you want to have the university closed during winter break in future years?*".

Stratum Number (h)	Employee Type	Population Size	No. of Surveys Distributed	Number of Responses	Number of Yes's	$\overline{y}_h$
1	Faculty	1374	500	232	167	0.72
2	Classified Staff	1960	653	514	459	0.89
3	Administrative Staff	252	74	67	58	0.87
4	Academic Professional	95	95	86	75	0.87
		3681	1322	899	759	

(a) [5 pts] Ignoring the strata and treating this as an SRS without replacement of size 899, calculate $\overline{y}_{srs}$ and $\sqrt{\text{Var}\left(\overline{y}_{srs}\right)}$, the SRS estimate and SE of the proportion of faculty and staff that responded "Yes". *HINT: Use the formula* $\hat{p}(1-\hat{p})$ *where* $\hat{p} = \overline{y}_{srs}$, *in your variance calculation.*

[Continued on next page...]

(b) [9 pts] Now treat this as a pre-stratified sample, and calculate $\overline{y}_{st}$ and $\sqrt{\text{Var}\,(\overline{y}_{st})}$, the stratified estimate and SE of the proportion of faculty and staff that responded "Yes". *[HINT: Use the number of responses as the sample size within each stratum; and use the same idea as in part (2a) for the needed variance calculations.]*

(c) [4 pts] Calcuate the DEFF for the stratified design and comment on whether it was worthwhile to stratify.

(d) [6 pts] Clearly not everyone who received a survey responded.

- Calculate the response rates in each stratum.
- Would you expect the bias due to nonresponse to be greater or less for the SRS estimates in part (2a), vs the stratified estimates in part (2b)?

3. [18 pts] The city council of a suburb of Chicago wants to know the proportion of eligible voters that oppose having a Chicago garbage incinerator opened in that suburb. They randomly select 100 residential phone numbers from the suburb's telephone book (which contains 3,000 residential numbers in all). Each selected residence is then called and asked for (a) the total number of eligible voters in that household and (b) the number of voters in the household opposed to the incinerator. A total of 157 voters were surveyed; of these, 23 refused to answer the question. Of the remaining 134 voters, 113 opposed to the incinerator, so the council estimates the proportion opposed as

$$\hat{p} = 112/134 = 0.83582$$

with

$$\text{Var}\,(\hat{p}) = 0.83582(1 - 0.83582)/134 = 0.00102$$

   (a) [6 pts] What is the **target population**? What is the **sampling frame**? *You do not have to use all the space provided.*

[Continued on next page. . .]

(b) [6 pts] A statistics intern working for the city council argues that this is a cluster sample. Interpreting this as a cluster sample,

- The psu's (primary sampling units) are the ________________________ .

- The ssu's (secondary sampling units) are the ________________________ .

(c) [6 pts] Are the estimates $\hat{p}$ and $\mathrm{Var}(\hat{p})$ given above valid? Why or why not? *(you do not have to use all the space provided)*

4. [18 pts] You are examining the data entered on an item measuring the number of miles driven per year in a travel survey about automobile trips. You encounter three cases from the total sample of 20,000:

   - One case reported 17,500 miles driven, but 1750 was entered
   - One case reported 17,599 miles driven, but 17,588 was entered
   - One case reported 17,599 miles driven, but 15,799 was entered

   In a separate analysis, it was found that the mean number of miles driven should be around 15,000.

   Give at least one strength and one weakness of each of the following strategies for identifying and correcting such coding errors.

   - [6 pts] Examining the distribution of the data as it was entered, and identifying and eliminating any outliers.
     - Strength:
     - Weakness:

[Continued on next page. . .]

- [6 pts] Drawing a 10% sample of the 20,000 responses, and identifying and correcting any errors found.
  - Strength:
  - Weakness:
- [6 pts] Examining all 20,000 responses and correcting any errors found.
  - Strength:
  - Weakness:

5. [20 pts] *Imputation methods.*

   (a) One method of imputation for missing responses to individual survey items is *mean imputation*.

      i. [4 pts] Explain briefly how mean imputation works.

      ii. [3 pts] Under what assumption (MCAR, MAR, MNAR) is mean imputation OK? (Choose one, no explanation needed.)

      iii. [3 pts] Identify a possible problem with mean imputation.

[continued on next page]

(b) Another method of imputation for missing responses is *regression imputation*.

i. [4 pts] Explain briefly how regression imputation works.

ii. [3 pts] Under what assumption (MCAR, MAR, MNAR) is hot-deck imputation OK? (Choose one, no explanation needed.)

iii. [3 pts] Identify a possible problem with regression imputation.